Segmenting the Earth: Challenges in Land Cover Classification

Shirley Cheng
scheng3@stanford.edu

Abstract

This project investigates the performance of U-Net-based models on the DeepGlobe Land Cover Classification dataset, evaluating the impact of three key modifications: focal loss, class-aware data augmentation, and transfer learning with a ResNet-34 encoder. Despite the widespread belief that class imbalance is the primary limitation in land cover segmentation, results suggest otherwise. Confusion matrices reveal that dominant classes like Agriculture are frequently misclassified as rare classes, leading to false positives rather than false negatives. This behavior is attributed to a combination of inter-class visual similarity, intra-class diversity, and a lack of global semantic context. Among the approaches tested, transfer learning with a pretrained encoder yielded the most consistent improvements, while focal loss worsened performance by amplifying false positives of rare classes. These findings highlight the need for models that better capture spatial context and global scene structure, suggesting future directions such as transformer-based architectures or satellite-specific pretraining.

1. Introduction

Land cover classification from satellite imagery is a crucial task in remote sensing with applications in urban planning, climate change analysis, and agricultural monitoring. This project tackles the Land Cover Classification task from the DeepGlobe 2018 Challenge [3], which presents a rich dataset of high-resolution satellite images labeled with seven semantic classes: urban, agriculture, rangeland, forest, water, barren land, and unknown (e.g., clouds or noise).

The input is a 1024×1024 RGB satellite image, and the objective is to produce an output of a segmentation mask where each pixel is assigned to one of six interpretable land cover classes (excluding unknown). Compared to other remote sensing datasets, the DeepGlobe dataset stands out for its diversity of land cover types and dense, high-quality annotations. The main challenges include inter-class similarity (e.g., rangeland vs. forest vs. agriculture), high intra-class variability (e.g., different types of crops in Agricultural land), and class imbalance.

To establish a baseline, I implemented a U-Net architecture trained from scratch using a class-weighted cross-entropy loss. I then explore how various techniques, such as focal loss, class-aware data augmentation, and transfer learning with ResNet-34.

This goal of this project is to investigate how these enhancements address the unique challenges of the DeepGlobe dataset, especially in terms of improving performance on rare or ambiguous land cover classes.

Despite common assumptions that class imbalance is the main limiting factor, my results suggest that inter-class similarity and intra-class variation—especially in ambiguous or visually similar regions—play a more significant role in limiting segmentation performance. The best-performing model (U-Net with ResNet-34 encoder) achieved a mean IoU of 0.34, improving over the baseline by 6.5 percentage points. These findings highlight the importance of global context and semantic reasoning in resolving ambiguous boundaries in satellite imagery.

2. Related Work

Land cover classification from satellite imagery is an important task in remote sensing. Traditional methods typically relied on classical machine learning algorithms such as Support Vector Machines (SVMs) [11] and Random Forests (RFs) [2]. While interpretable, these models often underperformed on complex scenes due to their limited ability to capture spatial and contextual information.

The emergence of deep learning, particularly convolutional neural networks (CNNs), has led to significant advances in semantic segmentation for remote sensing applications. Long et al. [9] first introduced the idea of that Fully Convolutional Networks (FCNs) [9] could be applied for dense, pixel-wise segmentation tasks, and not just image classification. This was further extended by architectures like U-Net [13], which introduced skip connections to better recover spatial detail during upsampling. Variants of U-Net have since become a staple in remote sensing applications, as demonstrated by Sherrah [14] and Audebert et al. [1].

A major challenge in remote sensing classification problems is that of class imbalance due to the inherent un-

equal distribution of real-world land types. Focal loss, first introduced by Lin et al.[8] for dense object detection, addresses imbalanced datasets by down-weighting well-classified examples and focusing training on hard, misclassified instances. Doi and Iwasaki[4] later adapted focal loss for high-resolution aerial image segmentation, demonstrating measurable gains—particularly for underrepresented classes.

While this project focuses on CNN-based methods, Transformer-based models have also recently gained traction and are now state-of-art in remote sensing segmentation tasks due to their ability to model long-range dependencies. SegFormer [17] introduced an efficient design that avoids heavy upsampling and has been adapted for overhead imagery. For example, recently, VistaFormer [10] proposed a scalable, lightweight transformer architecture specifically for satellite image time series segmentation. By replacing traditional Multi-Head Self-Attention with Neighbourhood Attention and using position-free encoding, VistaFormer achieves state-of-the-art accuracy on semantic segmentation of remote-sensing images, while using significantly fewer parameters and computational resources.

Overall, the field has progressed from classical models to CNN-based methods and now toward Transformers. My project will build on the CNN foundation, evaluating how enhancements like loss functions, data augmentation, and pretrained encoders affect segmentation performance.

3. Methods

I examine the impact of several modeling and training choices on land cover segmentation performance. My baseline is a U-Net architecture trained from scratch using a weighted cross-entropy loss. I then conduct an ablation study to investigate the effectiveness of three design decisions:

- Using focal loss instead of weighted cross-entropy to mitigate class imbalance.
- Applying class-aware data augmentation to increase model robustness and class diversity.
- Incorporating a ResNet-34 encoder pretrained on ImageNet to leverage transfer learning.

Each of these components is evaluated in isolation relative to the baseline(for the sake of time and scope of the project, I do not explore interactions between factors).

3.1. Baseline: U-Net Architecture

My baseline model is a U-Net architecture that I implemented from scratch. It follows a symmetric encoder-decoder structure with skip connections between corresponding layers. The encoder consists of four blocks, each containing two convolutional layers followed by ReLU activations and a 2×2 max-pooling operation for spatial downsampling. The decoder mirrors this structure, using transposed convolutions for upsampling, followed by two convolutional layers and skip connections that concatenate encoder features to preserve spatial detail. The bottleneck block connects the encoder and decoder at the lowest resolution, and a final 1×1 convolution projects the decoder output to class logits.

Given an input image $X \in \mathbb{R}^{3\times H\times W}$, the model outputs pixel-wise class scores $\hat{Y} \in \mathbb{R}^{C\times H\times W}$ for $C = 7$ land cover classes:

$$\hat{Y} = f_{\text{U-Net}}(X)$$

Following the practice suggested by the original DeepGlobe authors [3], I excluded the *Unknown* class from training and evaluation. This class typically represents cloud cover or other obscured regions, which are infeasible to classify and can introduce noise into the learning signal. In my implementation, I ignore this class by setting its weight to zero and ignoring it in the loss function.

3.2. Loss Function: Weighted Cross-Entropy vs Focal Loss

A key challenge in land cover segmentation is class imbalance, where certain classes (e.g., agricultural land) dominate the pixel distribution while others (e.g., water bodies) are rare. The exact distribution of class pixels across the training dataset is shown in Table 1.

To address this, I use weighted cross-entropy loss in the baseline, where class weights are computed using median frequency balancing:

$$w_c = \frac{\text{median}(f)}{f_c}$$

where f_c is the frequency of class c (excluding the ignored class). The loss becomes:

$$\mathcal{L}_{\text{WCE}} = -\sum_i w_{y_i} \log p(y_i)$$

As an ablation factor, I explore using focal loss [4], which modifies the standard cross-entropy loss to focus learning on difficult or minority-class pixels. Focal loss is expressed as:

$$\mathcal{L}_{\text{Focal}} = -\sum_i w_{y_i}(1 - p(y_i))^\gamma \log p(y_i)$$

where $p(y_i)$ is the model's predicted probability for the true class y_i, w_{y_i} is a class-specific weight, and $\gamma = 2.0$ is a focusing parameter.

Focal loss works by reducing the relative loss contribution from well-classified (easy) examples, where $p(y_i)$ is

high, and instead amplifies the gradient signal from misclassified (hard) examples. This makes it particularly effective for imbalanced datasets, where focal loss allows the model to concentrate more on learning the minority classes that are typically harder to classify correctly.

3.3. Data Augmentation: Class-Aware Oversampling and Transformations

I also explore whether data augmentation can enhance generalization and improve the model's ability to segment rare classes. While Shorten and Khoshgoftaar [15] suggest oversampling rare classes until parity is achieved, a key challenge in the context of land cover classification is that one image usually contain multiple classes.

To address this, I built a custom dataset wrapper that adaptively oversamples training images based on the presence of rare or very rare classes. For each training mask, the dataset expansion logic is as follows:

- If the image contains any **very rare class** (defined as Water), it is repeated up to 6 times.
- If it contains a **rare class** (defined as Rangeland or Forest), it is repeated 3 times.
- If the image is dominated by **Agricultural Land** (more than 70% of pixels) and does not contain a rare class or very rare class, it is excluded with a skip probability of 0.7.
- Otherwise images with only Urban, Rangeland, Forest land are repeated 2 times to maintain diversity.

For all repeated samples (except the first occurrence), I applied data augmentations to increase variability. These include horizontal and vertical flips, random rotations within $\pm 30^{\circ}$, and jitter. This targeted oversampling and augmentation strategy aims to improve the model's exposure to underrepresented classes during training.

In practice, however, the class aware augmentation only lead to a moderate improvement in representation of rare classes, as can be seen in Table 1. A key challenge seems to be that most images with rare classes also have a dominant class present. As thus, future work could explore more advanced rebalancing techniques. Nevertheless, since the sampling strategy still increases the amount of training data on the whole, it may still be insightful to see if even a moderate rebalancing is sufficient to impact model training dynamics.

3.4. Transfer Learning: U-Net with ResNet-34 Encoder

I evaluate the effect of transfer learning by replacing the baseline U-Net encoder with a ResNet-34 [6] backbone pretrained on ImageNet. Due to the small size of the training dataset(only 643 images), this approach may improve segmentation and accerlate training by leveraging pre-trained visual features.

Class Name	Original (%)	Augmented%
Urban land	11.04	10.53
Agriculture land	57.35	52.01
Rangeland	8.26	8.63
Forest land	11.54	10.67
Water	3.11	3.86
Barren land	8.64	9.21
Unknown	0.06	0.07

Table 1. Class-wise pixel percentage distribution in the training dataset before and after augmentation.

The decoder mirrors the U-Net structure and integrates skip connections from each ResNet block. The architecture outputs logits of shape $\mathbb{R}^{C \times H \times W}$, consistent with the baseline for a fair comparison.

3.5. Software

All models were implemented using PyTorch [12] with pretrained encoders from Torchvision. Data processing and evaluation made use of NumPy [5], Matplotlib [7], and Seaborn [16].

4. Dataset

The DeepGlobe Land Cover Classification dataset consists of 1,146 high-resolution satellite images at a resolution of 2448×2448 pixels [3]. Due to hardware and memory constraints, all images were resized to 1024×1024 which is sufficient to retain meaningful semantic detail. Of the 1,146 images provided by DeepGlobe, only the 803 training images include labeled RGB segmentation masks; the validation and test labels are withheld for competition evaluation. Therefore, for this project, I used only the labeled training set and manually split it into training, validation, and test sets with an 80/10/10 ratio, resulting in 643 training images, 80 validation, and test 80 images.

During preprocessing, I normalized all images to [0,1]. I also converted the RGB mask values into integer class labels (0 through 6), corresponding to the seven defined land cover categories. Data augmentation, described in detail in the Methods section, is included as an ablation factor to assess its impact on generalization and rare class performance.

A representative example of the expected model input and corresponding output is shown in Figure 1, where the raw satellite imagery is paired with a ground truth segmentation mask.

For a description of each class and its corresponding color in the segmentation mask, refer to Table 2

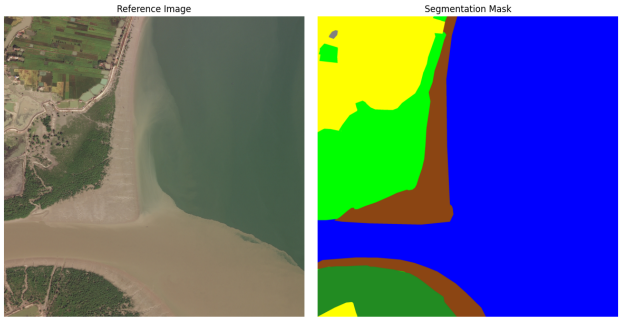

Figure 1. An example input-output pair from the dataset.

Class Name	Color	Description
Urban Land	Gray	Man-made, built-up areas.
Agricultural Land	Yellow	Farms, cropland, orchards, vineyards, nurseries
Rangeland	Light Green	Non-forest, non-farm green land covered with grass or sparse vegetation.
Forest Land	Dark green	Land with a certain percentage of tree density.
Water	Blue	Rivers, oceans, lakes, wetlands, and ponds.
Barren Land	Brow	Areas with little to no vegetation, including mountains, rocky terrain, deserts, and beaches.
Unknown	Black	Areas obscured by clouds or other phenomena making classification infeasible.

Table 2. Land cover class definitions with corresponding color labels and descriptions.

5. Experiments, Results, and Discussion

5.1. Training Details

I trained all models using the Adam optimizer with a learning rate of 1×10^{-4}, a batch size of 2, and for 30 epochs. The batch size was limited to 2 due to GPU memory constraints. For the learning rate, I experimented with both 1×10^{-3} and 1×10^{-4} on the baseline model; I found that the higher rate led to unstable training and spiky validation loss, while 1×10^{-4} provided more stable and consistent convergence. To determine the number of training epochs, I experimented using 10, 20, and 30 epochs using the baseline model and found that 30 was necessary to reach convergence when using the lower learning rate.

I chose the Adam optimizer because it combines the benefits of momentum and adaptive learning rates, making it well-suited for training segmentation networks like U-Net without extensive learning rate tuning. Its ability to handle sparse gradients and noisy updates was beneficial given the complexity and imbalance of the dataset.

The main evaluation metric is pixel-wise Intersection over Union, where Intersection over Union for class j is defined as:

$$\text{IoU}_j = \frac{\sum_{i=1}^{n} TP_{ij}}{\sum_{i=1}^{n} TP_{ij} + \sum_{i=1}^{n} FP_{ij} + \sum_{i=1}^{n} FN_{ij}} \quad (1)$$

Where:

- n is the total number of images in the dataset.
- k is the total number of land cover classes.
- TP_{ij} is the number of pixels in image i that are correctly predicted as class j (true positives).
- FP_{ij} is the number of pixels in image i that are incorrectly predicted as class j (false positives).
- FN_{ij} is the number of pixels in image i that belong to class j but are predicted as another class (false negatives).

Mean IOU would be computed as such:

$$\text{mIoU} = \frac{1}{k} \sum_{j=1}^{k} \text{IoU}_j \quad (2)$$

Table 3 summarizes the results of each model variant. Additionally, I also report IoU per class in Table 4. In the following subsections, I analyze the contribution of each component in detail.

Model Variant	mIoU
Baseline (U-Net)	0.27
+ Focal Loss	0.20
+ Data Augmentation	0.29
+ ResNet-34 Encoder	0.34

Table 3. Ablation study showing the effect of each individual modification on mean Intersection-over-Union (mIoU).

Class	Baseline	+Focal Loss	+Augmented	+ResNet-34
Urban	0.46	0.35	0.40	0.58
Agriculture	0.56	0.59	0.55	0.63
Rangeland	0.12	0.03	0.15	0.16
Forest	0.23	0.17	0.27	0.25
Water	0.27	0.16	0.30	0.34
Barren	0.10	0.07	0.15	0.18

Table 4. Ablation study showing the effect of each individual modification on **classwise** IoU.

5.2. Baseline Model

Overall, performance on baseline model was rather poor, with 0.27 mIOU. Examining class wise IoU, shown in Figure 2 reveals that while the model is able to learn representations for the dominant class of Agriculture, it struggles significantly with learning rare classes.

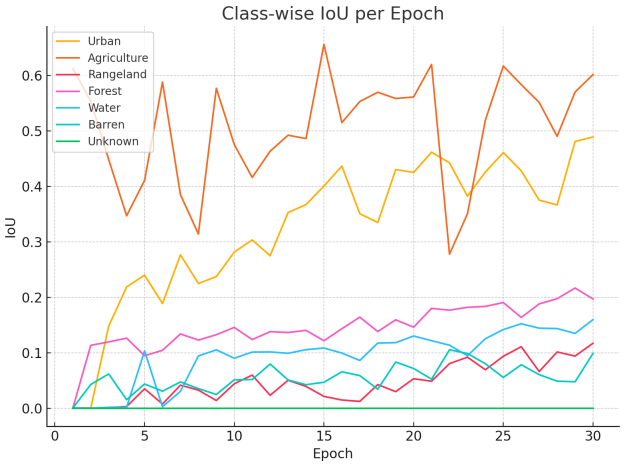

Figure 2. Per-class IoU of validation set across training epochs when using baseline model

Interestingly, although Urban Land makes up a relatively small proportion of the training data, the model is still able to learn representations for this class. This may be due to the visual distinctiveness of urban areas compared to other rare classes. In contrast, the model struggles to learn rangeland and Barren land which are more visually ambiguous and may exhibit overlaps with Agriculture. Water also has relatively poor performance. This may be due to its low presence in the dataset and its relative ambiguity.While we might expect Water to be highly distinctive, in practice, water in satellite imagery often appears as a dark, flat region that can resemble shadowed terrain or soil, rather than the clear, reflective blue we conventionally associate with it. This can be seen in the qualitative predictions shown in Figure 3.

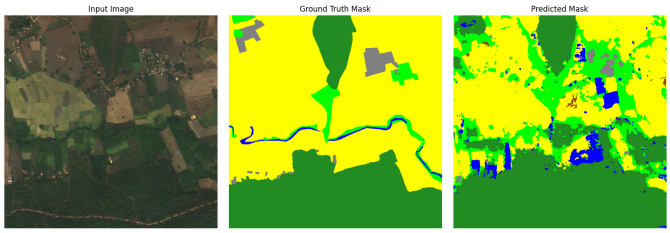

Figure 3. Input image(right), Ground Truth Mask(middle), Predicted Mask(left) for Baseline

The baseline model demonstrates a strong ability to capture fine-grained spatial detail, but often at the expense of semantic accuracy. As shown in Figure 3, the model predicts small patches of Rangeland (light green) within a larger Agriculture region, corresponding to minor vegetation patches visible in the input image. Impressively, it even detects isolated trees, visible as small dark green dots in the predicted mask, which align with tree locations in the input image. However, the model fails to incorporate broader contextual cues: these vegetation patches and trees are part of a cultivated field and should therefore be classified as Agriculture. The model's predictions often align with fine structures and edges in the input image, but this over-reliance on local texture and appearance leads to semantically incorrect segmentation. This highlights a key limitation of the baseline U-Net: its inability to integrate local detail into a coherent global understanding of the scene.

Finally, while the presence of class imbalance might suggest a tendency to overpredict the dominant class, our qualitative example reveals otherwise(Figure 3). We do not consistently see Agriculture replacing other classes; instead, Agricultural regions are frequently misclassified as other land types. This is evident in the confusion matrix(Figure 4, where false positives are disproportionately associated with rare classes being predicted in place of Agriculture. One possible explanation for this behavior lies in the high variability within the Agriculture class–driven by differences in crop type, growth stage, and field structure–which makes it difficult for the model to learn a consistent visual representation.

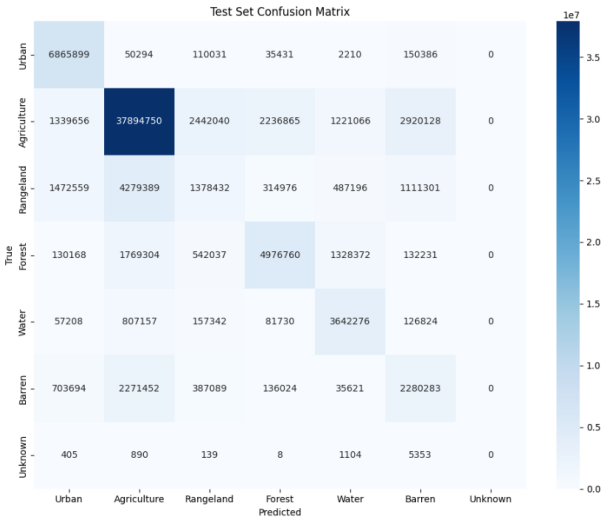

Figure 4. Confusion matrix for the baseline model. Most classes are not overpredicted as Agriculture, except Rangeland and Barren Land (which are visually similar). Agricultural land is often misclassified as rare classes, indicating a tendency toward false positives of rare classes.

5.3. + Focal Loss

To assess whether focal loss could improve performance on rare or underrepresented classes, I replaced the weighted cross-entropy loss in the baseline model with focal loss. This formulation was motivated by its success in object detection settings with severe class imbalance [8]. However, the model actually showed worsened performance on rare

classes (Table 4), indicating that focal loss does not help–and may even hinder–rare class learning in this context.

One possible explanation for this degradation is the relatively small size of the training set: there may simply not be enough high-quality, informative examples from rare classes to guide learning effectively. Moreover, the model was already prone to false positives for rare classes, as discussed earlier, with classes like Water and Rangeland frequently predicted in place of Agriculture. Since focal loss explicitly upweights hard-to-classify instances, it may have inadvertently amplified this existing tendency, especially in the presence of label noise or ambiguous pixels. For example, if some agricultural regions contain dark or flat patches resembling Water, and a few of these are mislabeled (or simply difficult), focal loss will push the model to fit these rare, hard examples more aggressively—leading to overprediction of the rare class.

This hypothesis is supported by the confusion matrix in Figure 5, which shows a substantial increase in Agricultural land being incorrectly identified as Water.

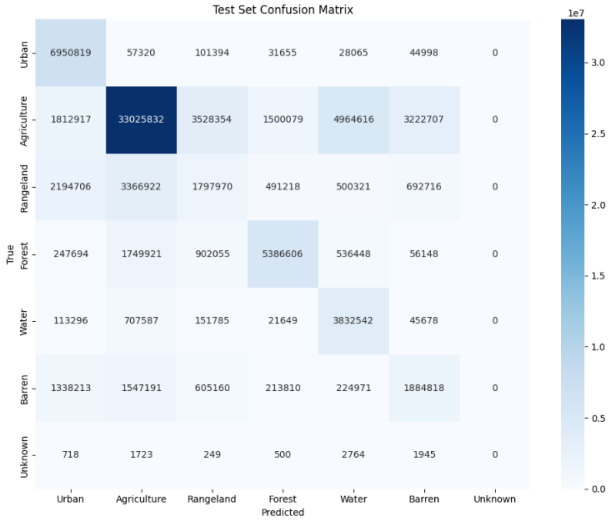

Figure 5. Confusion Matrix at Test Time(Focal loss). Compared to baseline, the amount of Agricultural land being incorrectly classified as Water increased.

These results suggest that while focal loss is theoretically well-suited for addressing class imbalance, its effectiveness may be limited in low-data regimes, particularly when visual ambiguity and noisy labels are already causing misclassifications in favor of rare classes.

5.4. + Data Augmentation

Data augmentation led to a slight overall improvement in mIoU, with most gains observed in underrepresented classes such as Forest, Water, and Barren Land (Table 4). However, it also resulted in decreased mIoU for dominant classes like Agriculture and Urban. This decline may possibly be attributed to the sampling strategy, which reduced the proportional frequency of these classes during training. Interestingly, despite undergoing a similar proportional reduction in training samples, Forest saw notable improvements in mIoU. This suggests that it may have benefited more from the diversity introduced through augmentation–especially color jitter. Forested regions often exhibit a wide range of natural color variation due to different tree types, seasons, and canopy structures. As such, color jitter likely enhanced the model's ability to generalize across intra-class diversity.

In contrast, classes like Agriculture and Rangeland were more negatively affected. These two classes are not only visually similar, but also both exhibit wide visual variations that may overlap with one another. For example, some types of rangeland may appear as neatly mowed grassland, while certain agricultural fields-especially during early growth stages-can resemble natural grasslands, making them difficult to distinguish. Applying color jitter in this case may have increased visual overlap between them, making them harder to distinguish. This is supported by the confusion matrix (Figure 6), which shows increased misclassification between Agriculture and Rangeland following augmentation. Additional qualitative examples of this confusion are shown in Figure 7.

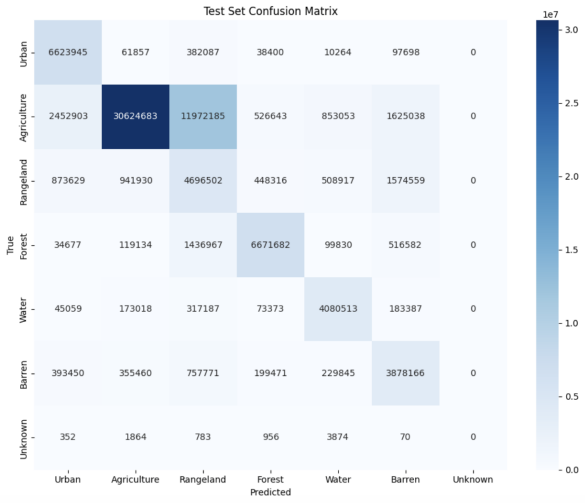

Figure 6. Confusion Matrix at Test Time(Augmentation). Agricultural land is frequently confused for Rangeland.

Overall, while data augmentation improved generalization and performance for several rare classes, it introduced new trade-offs, particularly for visually similar categories. Future work could explore more targeted augmentation strategies that preserve key inter-class distinctions–such as selectively applying jitter based on class characteristics–to boost minority class performance without degrading accuracy on more abundant or visually ambiguous classes.

Figure 7. Input image(right), Ground Truth Mask(middle), Predicted Mask(left) for Augmentation. Large patches of Agricultural land got confused for Rangeland.

5.5. + ResNet-34 Encoder

Replacing the baseline U-Net encoder with a ResNet-34 backbone pretrained on ImageNet improved performance across all classes, leading to an overall increase in mIoU (Table 4 and Table3).

The model was able to learn representation for non-dominant classes over training, with improvements in Water, Barren land, and Urban land, as shown in Figure 8.

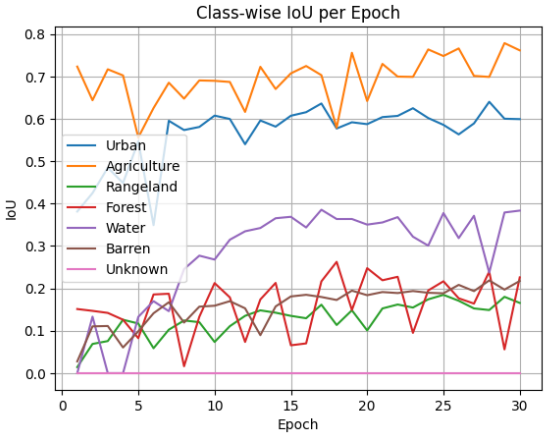

Figure 8. Per-class IoU of validation set across training epochs when using ResNet encoder

This improvement may be attributed to transfer learning: the pretrained encoder provides a stronger feature representation, especially in early layers,enabling the network to extract more meaningful semantic features from limited training data, leading to faster convergence (Figure 9).

Qualitative results also reflect the improvement in prediction quality. As shown in Figure 10, the ResNet-augmented model produces masks with fewer spurious predictions and more coherent predictions compared to the baseline. This suggests that the pretrained encoder helps the model better understand spatial context and reduce the noisy, fragmented patches seen in the base U-Net output.

Despite the overall improvement in segmentation quality, the model still struggles with mid-level semantic decisions, particularly in visually ambiguous regions. As illustrated in Figure 10, large portions of the agricultural field of

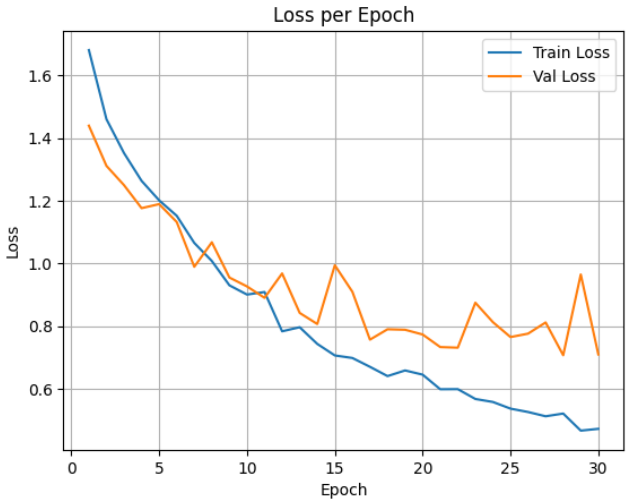

Figure 9. Training and Validation Loss using ResNet Encoder, demonstrates convergence roughly around the 15th epoch

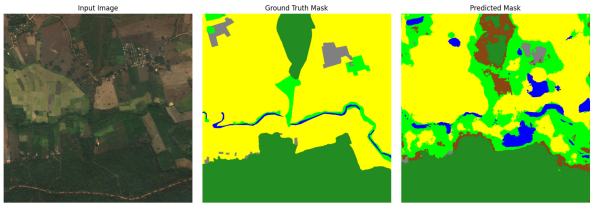

Figure 10. Input image(right), Ground Truth Mask(middle), Predicted Mask(left) for + ResNet-34.

the input image are misclassified as Rangeland(two visually similar classes) in the predicted mask. These errors suggest that, despite improvements from the ResNet encoder, the model still lacks sufficient global semantic reasoning to recognize large, coherent land cover structures. It struggles to consistently identify entire regions of Agriculture or Forest, instead segmenting them into smaller, disjointed parts. Addressing this may require architectural changes that support scene-level understanding, such as attention mechanisms or hierarchical context modeling.

6. Conclusions

This project explored the challenges of semantic segmentation in high-resolution satellite imagery for land cover classification. Overall, the task was difficult, with all U-Net model variants achieving relatively low performance, with mIoU scores ranging from 0.20 to 0.34. While class imbalance is often cited as a key limitation in such tasks, the results here suggest it is not the primary bottleneck. Across the confusion matrices for all model variants, there is little evidence that rare classes are being misclassified as the dominant class, Agriculture. On the contrary, we observe the opposite pattern–particularly pronounced in models

trained with Focal Loss and data augmentation–where Agriculture is frequently misclassified as rare classes. Therefore, the low mIoU appears to result not from an inability to detect rare classes, but from a tendency to incorrectly assign rare class labels, leading to a high number of false positives of rare classes.

My hypothesis is that three factors collectively contribute to this behavior:

- Low-Data Regime: Limited training examples hinder the model's ability to learn the spatial extent and contextual cues associated with each class. As a result, the model tends to overgeneralize from sparse features, applying class labels too broadly across the image. This is particularly problematic in satellite imagery, where local features (e.g., patches of green) are shared by multiple classes and insufficient data makes it difficult to learn precise distinctions.
- Inter-Class Similarity: Classes such as Agriculture, Rangeland, and Barren Land often share similar color and texture profiles, especially after augmentations like color jitter. Their visual similarity makes it hard to learn distinct boundaries between classes, resulting in misplaced predictions.
- Intra-Class Variation: Classes exhibit wide internal diversity (e.g., different color and textures of vegetation), requiring the model to learn broad, abstract representations. Without strong semantic guidance, the model tends overgeneralize, labeling any region with matching low-level features, contributing further to false positives.

Another challenge stems from the spatial scale of satellite imagery. Each image spans approximately $1.5km^2$[3], encompassing numerous fine-grained visual patterns. The task is not just to recognize texture but to interpret small-scale visual cues within a larger semantic context. For instance, a small grassy patch within a farm should still be classified as Agriculture, not Rangeland. U-Net is highly capable of capturing fine details, but struggles to integrate them into broader structural understanding.

Among the techniques tested, transfer learning with a ResNet-34 encoder pretrained on ImageNet yielded the most substantial improvement. The pretrained model produced more spatially coherent predictions and demonstrated better local context aggregation, correctly associating visually ambiguous patches with their broader land use category. For example, it was more likely to correctly classify isolated grassy areas within a field as part of Agriculture, rather than as Rangeland. This suggests that mid-level features and boundary cues learned from natural images help the model interpret ambiguous satellite textures more effectively.

However, these improvements were localized rather than holistic. While the pretrained encoder helped resolve small-scale ambiguities, the model continued to struggle with large-scale decisions–sometimes misclassifying parts of agricultural fields as Rangeland. This indicates that the model still struggles to understand the larger global semantic structure, especially in the context of visually ambiguous classes.

Data augmentation also showed mixed results. The class-aware oversampling strategy improved performance for rare classes such as Forest, Water, and Barren Land, which benefited from increased diversity. However, it simultaneously introduced more confusion between Agriculture and Rangeland, reinforcing the idea that augmentation alone is not enough when classes are visually similar. Without explicit structural or contextual cues, augmentation may even amplify class overlap.

Focal loss, although intended to address class imbalance, actually worsened performance–highlighting that the core issue lies not in an imbalanced dataset, but in the ambiguity of class boundaries. In a setting where many hard examples are ambiguous rather than informative, placing extra emphasis on them may increase noise sensitivity. When class boundaries are subtle or poorly defined, penalizing low-confidence predictions can push the model to make over-confident, incorrect decisions, increasing false positives.

Taken together, these results support the conclusion that inter-class similarity and intra-class variation—amplified by limited data and lack of global context—are the primary barriers to effective land cover segmentation, not class imbalance alone. This highlights the need for models that can capture broader scene-level context to make more semantically informed predictions. At the local level, land cover classes often exhibit overlapping visual features, making isolated patches difficult to interpret reliably. Global context provides the surrounding semantic and spatial information needed to disambiguate such cases. For instance, recognizing that a patch of vegetation lies within the boundaries of a regular, rectangular field—surrounded by other cultivated areas—strongly suggests that it is part of Agriculture, even if its local texture varies. Similarly, identifying a water region as a coherent body (e.g., a river or lake) based on its shape, continuity, and relation to adjacent landforms can help confirm its class, even if the local appearance is dark or noisy. By reasoning over the structure and relationships of larger regions, models with access to global context may be able to make more semantically accurate predictions.While ResNet-based transfer learning helps mitigate these issues to some extent, further improvement will likely come from models capable of capturing long-range spatial dependencies and reasoning over the global structure of a scene. Potential future directions include transformer-based architectures, self-supervised pretraining on satellite-

specific data, and hierarchical models that integrate local detail with broader scene understanding.

References

[1] N. Audebert, B. Le Saux, and S. Lefèvre. Joint learning from earth observation and openstreetmap data to get faster better semantic maps. In *EARTHVISION 2017 IEEE/ISPRS CVPR Workshop. Large Scale Computer Vision for Remote Sensing Imagery*, Honolulu, United States, 2017.

[2] M. Belgiu and L. Drăguţ. Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114:24–31, 2016.

[3] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[4] K. Doi and A. Iwasaki. The effect of focal loss in semantic segmentation of high resolution aerial image. In *2018 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 6949–6952, Valencia, Spain, 2018. IEEE.

[5] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, et al. Array programming with numpy. *Nature*, 585(7825):357–362, 2020.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[7] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[8] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017.

[9] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *arXiv preprint arXiv:1411.4038*, 2015. Presented at CVPR 2015.

[10] E. MacDonald, D. Jacoby, and Y. Coady. VistaFormer: Scalable Vision Transformers for Satellite Image Time Series Segmentation. *arXiv preprint arXiv:2409.08461*, 2024.

[11] M. Pal and P. M. Mather. Support vector machines for classification in remote sensing. *International Journal of Remote Sensing*, 26(5):1007–1011, 2005.

[12] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32:8026–8037, 2019.

[13] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *arXiv preprint arXiv:1505.04597*, 2015. Conditionally accepted at MICCAI 2015.

[14] J. Sherrah. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. *arXiv preprint arXiv:1606.02585*, 2016.

[15] C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

[16] M. L. Waskom. Seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.

[17] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *arXiv preprint arXiv:2105.15203*, 2021.

7. Contributors

This is individual project, I worked everything on my own. This is not shared with another course.